Supplementary Material for: Diffusion pseudotime robustly reconstructs lineage branching

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, Fabian J. Theis

Contents

List of Figures

List of Tables

1 Actual time, universal time, pseudotime

Cell differentiation is a largely asynchronous process. Even if we consider a single-fated lineage, due to the stochastic nature of the system a heterogeneous population of cells coexist at any given time. Although each single cell takes a different trajectory in actual time (due to the stochasticity in the differential equation), all these trajectories lie on a common manifold in the genes expression space ($C \subset \mathbb{R}^G$), where G denotes the number of genes. The manifold C (if one dimensional) can be parametrized by the arc length s along C.

Let us study a single cell trajectory along the manifold. For this, we can assign a velocity $\boldsymbol{v}(t)$ to each time point t that is approximately tangent to the manifold C. (Supplementary Fig. 1). If we consider an equidistant temporal sampling of the single cell trajectory, the tangent velocity is inversely proportional to the density $\rho(t)$ of the cell states on the trajectory at that time point, that is $|\boldsymbol{v}(t)| = 1/\rho(t)$. In other words the more the time points of the single cell trajectory happen to be in a region of $\mathbb{R}^G$ (black circle in Supplementary Fig. 1B), the slower the single cell has passed through that region. Because $\boldsymbol{v}(t)$ is tangent on C we can write

$$\mathrm{d}s = |\boldsymbol{v}(t)|\mathrm{d}t = \frac{1}{\rho(t)}\mathrm{d}t. \tag{1}$$

Integrating ds, starting at the root cell, along C up to actual time t yields the arc length, which we refer to as *universal time*

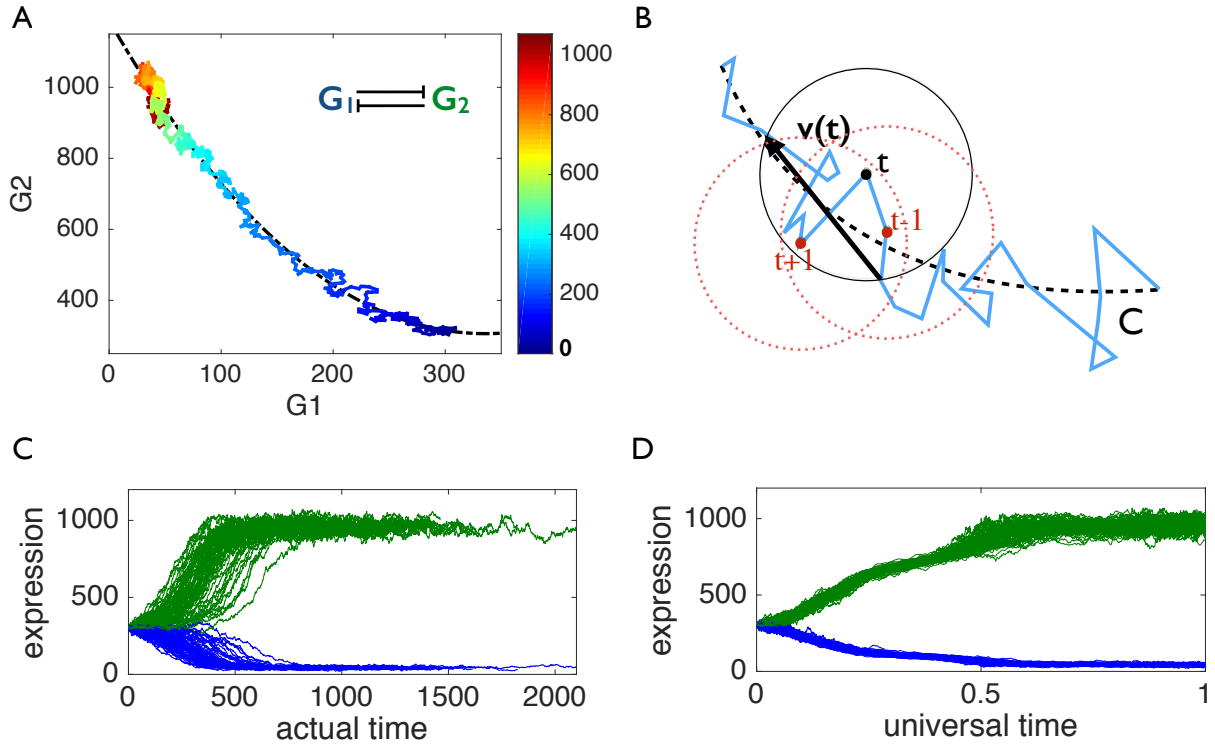
$$s(t) = \int_{C:[s(0),s(t)]} \mathrm{d}s = \int_0^t |\boldsymbol{v}(t')|\mathrm{d}t' = \int_0^t \frac{1}{\rho(t')}\mathrm{d}t'. \tag{2}$$

This assigns a universal time $s(t)$ to every actual time single cell trajectory as measured in time lapse microscopy. However, for snapshot data it is often difficult to learn the original manifold C and obtain universal time.

In the context of snapshot data, the common practice is to first map the data to a new space where noise is diminished and thereby the manifold becomes more pronounced. Then in general one can define pseudotime as the distance (arc length) to the root cell on the mapped manifold (call it C'). Such notion of pseudotime as an arc length has also previously been used in [1, 2, 3] and [4]. For diffusion pseudotime, the mapping is from $\mathbb{R}^G$ to $\mathbb{R}^{n-1}$ where n is the number of cells, and distances on C' are characterized by a new metric we term "diffusion pseudotime" (see following sections and main text). Supplementary Figure 2 illustrates the three concepts (actual time, universal time, and pseudotime) and how they are related to each other. Thus we established a unified framework which can be used to bring time-laps microscopy data and single-cell snapshot expression data together and comparable (e.g. data [5]). The connection of universal time with pseudotime as established here is valid as long as several cells from several developmental parts are present in the snapshot sampling. But we don't make any assumption of stationary state sampling (as used in [3]). This is especially helpful in the context of single-cell snapshot data where sampling densities are usually far from any stationary state and influenced by cell division rates and noise and are very subjective to the experimental design.

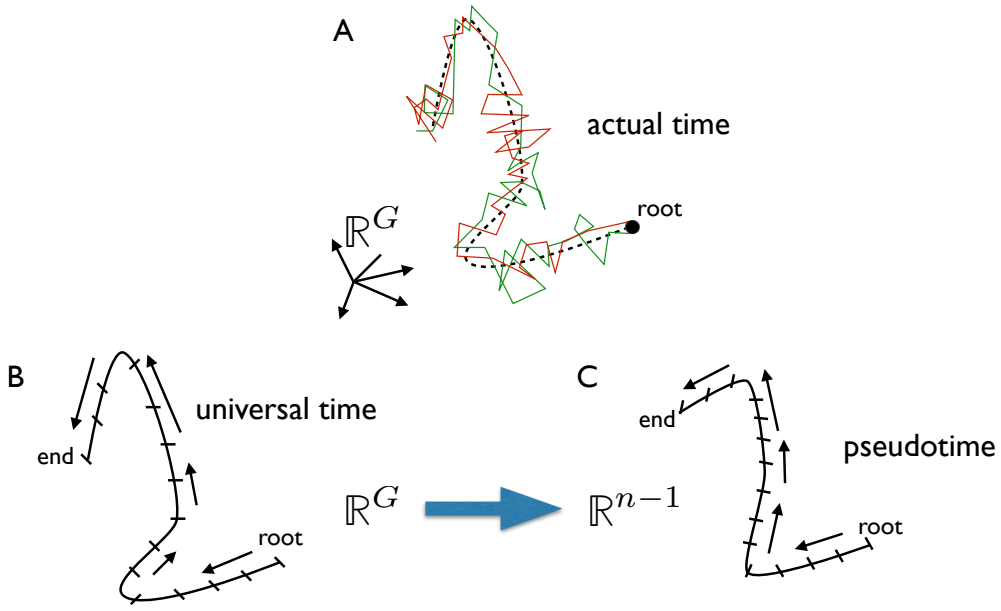
As a demonstration for asynchrony among single cells (even though from a single lineage), we simulated 100 cells (with a gene regulatory network of 6 genes [6]) starting from the same state at time zero (Supplementary Fig. 3A). We show that, when plotting versus universal time, all expression trajectories of these asynchronous single cells are brought to a unique expression curve that is the universal gene expression trajectory (Supplementary Fig. 3B) for that lineage.

Supplementary Figure 3D shows the Wanderlust pseudotime for the toy data and Figures 3 E) to G) show diffusion pseudotime on several mappings C' depending on the choice of the used diffusion map method and its respective parameter (see caption of Supplementary Fig.3) .

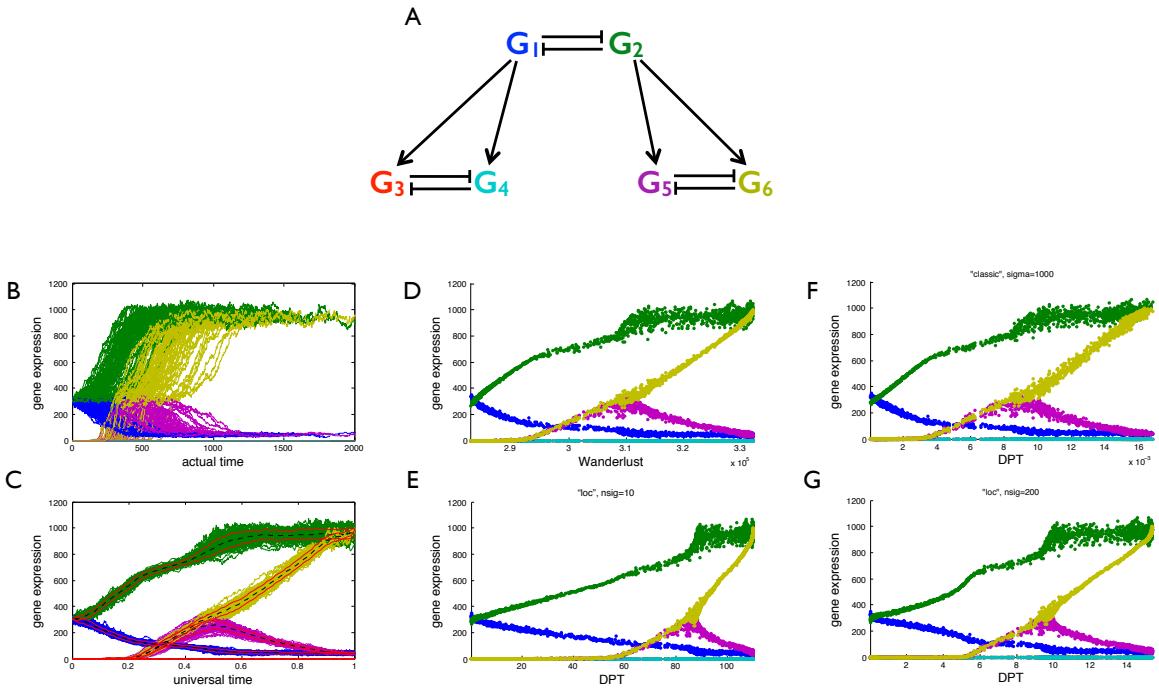


Supplementary Figure 1: A) A single cell trajectory for a toggle switch.The color indicates actual time. The trajectory is adjacent to a manifold C (dashed line) which we hypothesis is the same for all single cell trajectories following the same dynamics model. B) For any single cell trajectory with equidistant sampling of time steps, a velocity $\boldsymbol{v}(t)$ is defined for each time point t that is approximately tangent to the manifold C and $|\boldsymbol{v}(t)|$ is approximately equal to the diameter of a circle centered at time point t divided by the number of time points inside the circle. That is $|\boldsymbol{v}(t)|$ is inversely proportional to the density of trajectory points at t. For each single cell trajectory, we calculate the universal time as $\int_0^t |\boldsymbol{v}(t')|dt'$. C) Expression of G1(blue) and G2(green) for several single cell trajectories versus actual time, exhibits large asynchrony between the single cells because of the stochasticity in the dynamics model. D) The asynchronous trajectories in C fall on top of each other when plotted against universal time.

We mention that distances on manifolds have been discussed in several other publications, but only for snapshot data, and without realizing the connection to actual time trajectories as measured in time-lapse microscopy. In section 9.4.2, we discuss this in detail.



Supplementary Figure 2: A) Two (red and green) actual time single cell trajectories in gene expression space($\mathbb{R}^G$). Each jump on a trajectory happens in an (equidistant) unit of actual time. B) universal time is defined as arc length on the data manifold starting from the root. This manifold $C \subset \mathbb{R}^G$ remains the same for several single cell trajectories, as well as for a snapshot sample of single cells. C) pseudotime (in general) is defined as arc length on a more pronounced mapped manifold C' (with respect to noise). In case of diffusion pseudotime this mapping is from $\mathbb{R}^G$ to $\mathbb{R}^{(n-1)}$ where n is the number of sampled cells.



Supplementary Figure 3: A) A toy gene regulatory network with 6 genes B) actual time simulation of expression for the single-fated lineage for which G_2 and consecutively G_6 win the toggle-switch competitions C) expression time series versus universal time D) expression of snapshot sampled data vs. Wanderlust pseudotime E) expression of snapshot sampled data vs. Diffusion pseudotime (locally rescaled diffusion map, $\kappa = 10$) F) expression of snapshot sampled data vs. Diffusion pseudotime (classic diffusion map,$\sigma = 1000$) G) expression of snapshot sampled data vs. Diffusion pseudotime (locally rescaled diffusion map, $\kappa = 200$)

2 "Locally scaled" diffusion map

Our "classic" implementation of diffusion maps for single-cells snapshot data is fully presented in [6]. For the data analysis in the current manuscript, we extended the implementation of [6] to use a locally varying kernel width. In [6] we suggested an interpretation of the Gaussian kernel in terms of interfering wave functions. In other words the Gaussian kernel can be decomposed into its multiplicand wave functions. This interpretation turns out to be very useful when the diffusion wave function varies (because of varying noise models) at the position of each cell. In practice we assume the Gaussian kernel width is different for each cell, and is determined by the distance of each cell to its κth nearest neighbor we will have:

$$Y_{\mathbf{x}}(\boldsymbol{x'}) = \left(\frac{2}{\pi\sigma_{\mathbf{x}}^2}\right)^{1/4} \exp\left(-\frac{||\boldsymbol{x'}-\boldsymbol{x}||^2}{\sigma_{\mathbf{x}}^2}\right) \tag{3}$$

The normalization of $Y_{\mathbf{x}}(\boldsymbol{x'})$ is such that $\int_{-\infty}^{\infty} Y_{\mathbf{x}}^2(\boldsymbol{x'})\mathrm{d}\boldsymbol{x'} = 1$.

$$K(\mathbf{x},\mathbf{y}) = \int_{-\infty}^{\infty} Y_{\mathbf{x}}(\boldsymbol{x'})Y_{\mathbf{y}}(\boldsymbol{x'})\mathrm{d}\boldsymbol{x'} = \left(\frac{2\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}}{\sigma_{\mathbf{x}}^2+\sigma_{\mathbf{y}}^2}\right)^{1/2} \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2(\sigma_{\mathbf{x}}^2+\sigma_{\mathbf{y}}^2)}\right) \tag{4}$$

Using K we build the density normalized Markovian transition matrix T.

$$T_{\mathbf{xy}} = \frac{1}{\tilde{Z}(\boldsymbol{x})}\frac{K(\mathbf{x},\mathbf{y})}{Z(\boldsymbol{x})Z(\boldsymbol{y})} \quad , T_{\mathbf{xx}} = 0 \tag{5}$$

$$\tilde{Z}(\boldsymbol{x}) = \sum_{\mathbf{y}\in\boldsymbol{\Omega}/\mathbf{x}} \frac{K(\mathbf{x},\mathbf{y})}{Z(\boldsymbol{x})Z(\boldsymbol{y})} \tag{6}$$

$$Z(\boldsymbol{x}) = \sum_{\mathbf{y}\in\boldsymbol{\Omega}/\mathbf{x}} K(\mathbf{x},\mathbf{y}) \tag{7}$$

Let us call the sorted right eigenvectors of T corresponding to the largest eigenvalues $\psi_i, i = 0,1,..,n-1$. As described in [6], $\psi_1,\psi_2,..\psi_k$ with k chosen at a significant gap between the corresponding eigenvalues $\lambda_1,\lambda_2,...,\lambda_k$, provide a low dimensional map of the original data manifold which resides in the original high dimensional space of gene expression.

3 Diffusion pseudotime

The t'th power of the transition matrix T represents a random walk of length t on the data graph. The transition matrix enables us to simulate the time propagation of a wave function (or probability) that has been localized to some specific region of the graph (e.g. a wave function that resides on the pluripotent cells only) at time zero.
Noting that T is a right stochastic matrix (i.e. row normalized), the time evolution (propagation) of the probability density $f(t)$ is described by the graph Laplacian matrix $L = I - T$ as follows:

$$\Delta f(t) = f(t)(-L) \tag{8}$$

or in terms of T:

$$f(t) = f(t-1)T = f(0)T^t \tag{9}$$

To account for the asynchrony of differentiating cells present in a snapshot data, one may study the term $\sum_{t=1}^{\infty} f(t)$ which provides the (time independent) path integral for reaching each cell from $f(0)$:

$$\sum_{t=1}^{\infty} f(t) = f(0) \sum_{t=1}^{\infty} T^t \tag{10}$$

The sufficient constraint for the sum above to converge, is that all eigenvalues of T^t were smaller than one. However, a stationary state exists for $v(\infty)$ which means that T has an eigenvalue equal to 1 with the corresponding right eigenvector $\psi_0 = \mathbb{1}$ (and the left, eigenvector equals the sampling density at the position of each cell $\phi_0(x) = Z(x)$). This means the sum in equation 10 diverges. However the stationary state contains information only about cells' sampling density and not about the consecutive stages of temporal evolution (i.e. no pseudotime information). Thus we can reduce the stationary component of T and easily perform the sum in equation 10. We call the new matrix M:

$$M = \sum_{t=1}^{\infty} (T - \psi_0 \phi_0^T)^t \tag{11}$$

$$= (I - (T - \psi_0 \phi_0^T))^{-1} - I \tag{12}$$

M is in fact the projected (on $\psi_i \quad , \quad 0 < i \leq n-1$ subspace) path integral matrix and shares the same eigenvectors with T (except for ψ_0):

$$\begin{aligned} M(x,z) &= \sum_{t=1}^{\infty} \left(T(x,z) - \psi_0(x) \phi_0^T(z) \right)^t \\ &= \sum_{t=1}^{\infty} \sum_{i=1}^{n-1} \lambda_i^t \psi_i(x) \phi_i^T(z) \\ &= \sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_i} \psi_i(x) \phi_i^T(z) \end{aligned} \tag{13}$$

If $f(0)$ is chosen localized at cell x (i.e. if $f(0) = \delta(x)$), $f(0)M$ will be a row of M which we present by $M(x,.)$. Moreover, we consider $M(x,.)$ as the feature representation for cell x to define a new (time invariant) projected path integral distance measure $D_M(x,y)$ as follows

$$\mathrm{dpt}^2(x,y) = ||M(x,.) - M(y,.)||^2_{1/\phi_0} \tag{14}$$

$$\begin{aligned} &= \sum_z \frac{(M(x,z) - M(y,z))^2}{\phi_0(z)} \\ &= \sum_{i=1}^{n-1} \left(\frac{\lambda_i}{1-\lambda_i} \right)^2 (\psi_i(x) - \psi_i(y))^2 \end{aligned} \tag{15}$$

dpt is a proper (weighted L^2 norm) distance metric. It is quite robust to noise and yet does not utilize low-dimensional approximations usually applied for visualization (embedding). This can facilitate the application of several common mathematical methods which were not applicable to the noisy Euclidean distances in the original $\mathbb{R}^G$ gene expression space. Thus we use $\mathrm{dpt}(f(0),.)$ (i.e. distance from the

root cell) as the pseudotime measure. In case of ambiguity in the position of the root cell, one can specify multiple root cells and take average distance to the roots ($< \mathrm{dpt}(r, .) >_s$) as the pseudotime measure. As Equation 15 indicates, dpt is closely (up to a scaling factor) related to diffusion distance. Whereas diffusion distance is Euclidian distance in diffusion map space with $\lambda_i^t \psi_i$ coordinates, diffusion pseudotime is Euclidean distance in $\frac{\lambda_i}{1-\lambda_i}\psi_i$ coordinates.

The idea of looking at stochastic processes (random walks) occurring at all times (e.g. average first passage time [7], average commute time) for data clustering has been studied in several previous publications [8, 9]. However non of the previously studied measures provides a proper metric. Introduction of dpt as a proper metric on a mapped space however, alleviates definition of diffusion pseudotime which grows (almost) linearly with arc length on C'. Thus dpt only depends on the topology of the manifold and will not be affected by several branching events or heterogeneous sampling densities at several regions of the manifold.

4 Branch assignment

As a robust and biologically relevant (in the context of cell development) metric on the data manifold, dpt can be used to identify the cells at the tip of the branches of data. Let us consider a manifold with only three branches. Knowing the root cell r_1 we first identify r_2 that maximizes $\mathrm{dpt}(r_1, r_2)$. For any cell residing on the (direct) connecting path between r_1 and r_2, the triangle inequality holds at its lower bound, i.e. $\mathrm{dpt}(r_1, x) + \mathrm{dpt}(x, r_2)$ is equal to (or only slightly bigger than) $\mathrm{dpt}(r_1, r_2)$. It is only for cells residing on the third branch that $\mathrm{dpt}(r_1, x) + \mathrm{dpt}(x, r_2)$ becomes significantly bigger than $\mathrm{dpt}(r_1, r_2)$. Thus the third tip cell r_3 can be identified as the cell maximizing the sum of distances to r_1 and r_2. In short :

$$r_2 = \arg\max_x \mathrm{dpt}(r_1, x) \tag{16}$$

$$r_3 = \arg\max_x \Big(\mathrm{dpt}(r_1, x) + \mathrm{dpt}(x, r_2)\Big) \tag{17}$$

Now we can perform a pseudotime ordering where the initial probability $f(0)$ is chosen zero everywhere except at the tip of a branch (either r_1, r_2 and r_3). The ordering on every two branches will correlate with each other only on the third branch and anti-correlate on the two branches themselves (see Fig. 1a (3) in the main text). We use this property to find a cutoff x for each branch. More precisely to separate branch 1, we first do three independent orderings $O1, O2, O3$ with assigning r_1, r_2 and r_3 as the root of ordering correspondingly. Then based on Kendall-tau correlations we build a new measure of concordance between the $O2$ and $O3$ orderings from s_1 until x and their anti-concordance for the rest of cells:

$$\begin{aligned} K_{2,3}(x) = &\,\mathrm{Kendall.tau}(O2(r_1 : x), O3(r_1 : x)) \\ &- \mathrm{Kendall.tau}(O2(x+1 : end), O3(x+1 : end) \end{aligned} \tag{18}$$

Finally we find the cutoff x such that

$$x_{O1} = \arg\max_x \Big(K_{2,3}(x) - K_{2,3}(x-1)\Big) \tag{19}$$

Such finite difference optimization choice is to avoid influence of densities on where the cutoff should be. Note that we used this formulation to enhance clarity. The implementation to compute $K_{2,3}(x)$ uses a more efficient, recursive form: $K_{2,3}(x) = K_{2,3}(x) + \Delta K_{2,3}(x)$ and $x_{O1} = \arg\max_x \Big(\Delta K_{2,3}(x)\Big)$.

In case of unknown root cell, DPT can provide suggestions by finding three cells at the tip of three different branches of the data manifold. Randomly picking a cell z and assigning $r_1 = \arg\max_x \mathrm{dpt}(z, x)$

will propose r_1 as the first suggestion. Two other suggestions r_2 and r_3 are obtained following equations 16 and 17.

DPT also allows choosing more than one root cell. This feature is useful when root belongs to a population of stationary state cells with large variance in their expression state. DPT would then assign a diffusion pseudotime to the data points (i.e. cells) which is the average of the dpt distance to each root cell:

$$\mathrm{DPT}(x) =< \mathrm{dpt}(x, \mathrm{root}_i) >_{\mathrm{root}_i} \tag{20}$$

where root_i are the root cells chosen by the user.

5 Detecting transcriptional changes

Transcriptomic data sets on the single-cell level are usually accompanied by non-negligible levels of noise. Moreover, the heterogeneity of cell populations shown in bimodal expression patterns needs to be addressed. We employed a two-part, generalized linear model that allows to quantify the proportion of cells expressing a certain gene as well as the mean expression level, a modified Hurdle model [10]. Briefly, let Y_{ig} the gene expression level of gene g in cell i. Then, an indicator variable Z_{ig} determines, whether gene g is expressed in cell i and the expression level of gene g given it is expressed, is determined by normal distribution:

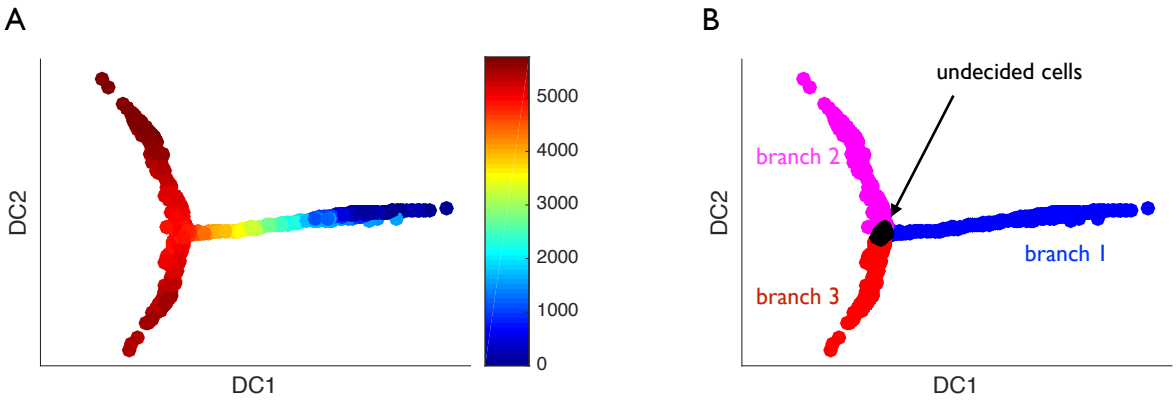
$$\begin{aligned}\mathrm{logit}\left(\mathrm{P}(Z_{ig}=1)\right) &= X_i\beta_g^D \\ \mathrm{P}(Y_{ig}=y|Z_{ig}=1) &= \mathcal{N}\left(X_i\beta_g^C, \sigma_g^2\right)\end{aligned}$$

We have two regression components, the discrete D and the continuous C component.
In order to compute a likelihood-ratio on two different populations, [10] developed a combined log-fold change defined as follows. For each gene g, let $u(x)$ the expected value of the continuous component, as $u(x) = \langle C, x\rangle$, and let $v(x) = (1 + \exp(-\langle D, x\rangle))^{-1}$ the expected value of the discrete component. The log-fold change from population p_1 to population p_2 is then defined as

$$lfc(p_1, p_2) = u(x|x \in p_2) \cdot v(x|x \in p_2) - u(x|x \in p_1) \cdot v(x|x \in p_1)$$

6 Toy data

For a six dimensional toy model generated by simulation of a stochastic toggle switch with downstream gene activation as shown in Supplelmentary Figure 3A, diffusion pseudotime is shown in Supplementary Figure 4A. DPT also robustly identifies samples in the decision region and the two downstream branches (Supplementary Fig. 4B).



Supplementary Figure 4: For toy model of differentiating and branching cells A) diffusion pseudotime is shown by color code on the diffusion map embedding of data. B) DPT robustly identifies samples in the decision region and the two downstream branches.

7 Early blood development data

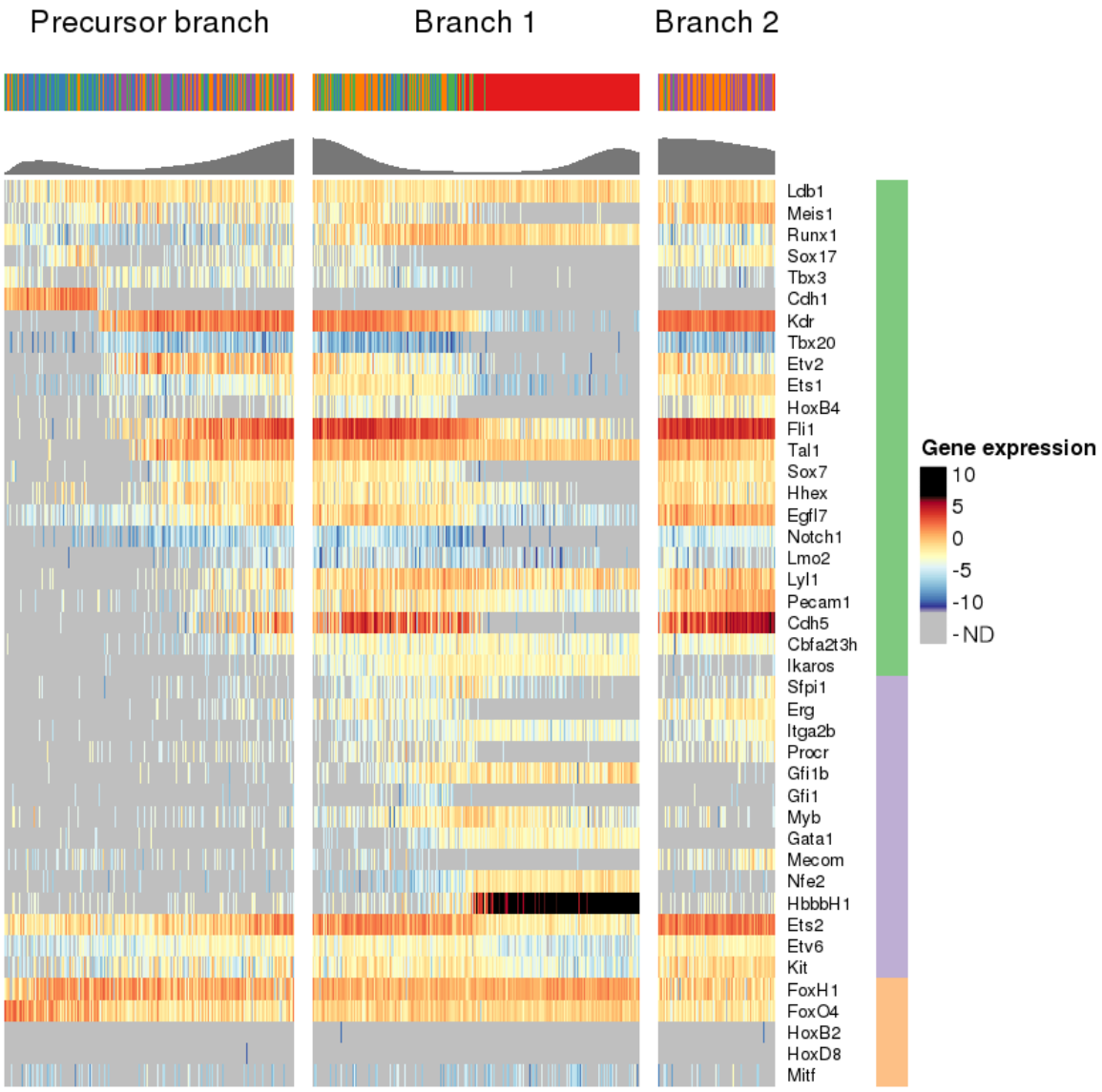
We reanalyzed a single-cell qPCR data set focusing on early blood development [11]. Data is publicly available in GEO with accession number GSE61470.

7.1 Pre-processing

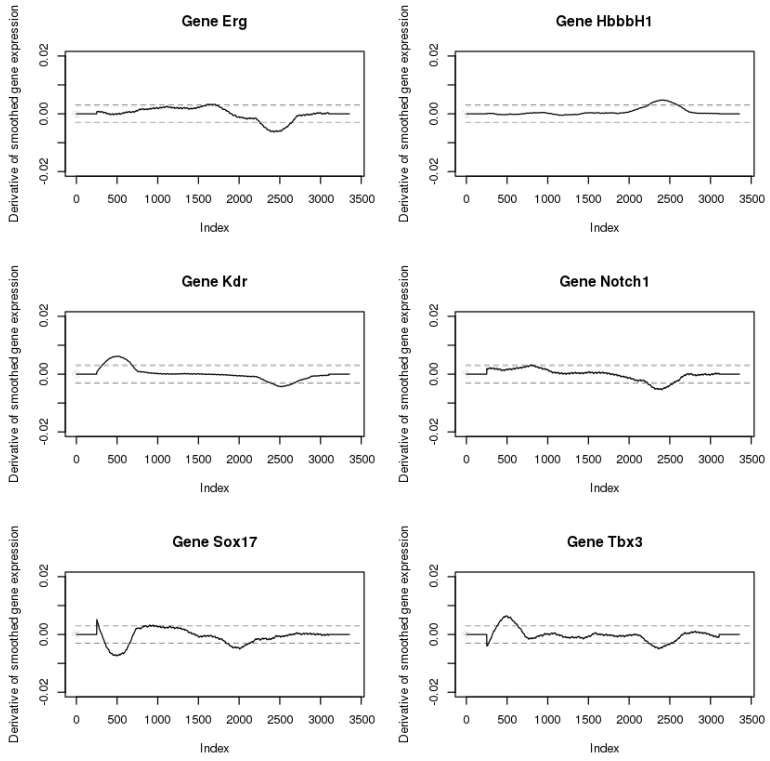
We reanalyzed a single-cell qPCR data set (normalized version with 3934 cells, 42 genes) focusing on early blood development [11]. Data is publicly available in GEO with accession number GSE61470. The authors of [11] suggest that normalization was done by subtraction of Gene expression from the limit of detection and normalization on a cell-wise basis to the mean expression of the four housekeeping genes (*Eif2b1, Mrpl19, Polr2a* and *Ubc*) in each cell. Cells that did not express all four housekeeping genes were excluded from subsequent analysis, as were cells for which the mean of the four housekeepers was ± 3 s.d. from the mean of all cells. A dCt value of 14 was then assigned where a gene was not detected.

7.2 Transition identification

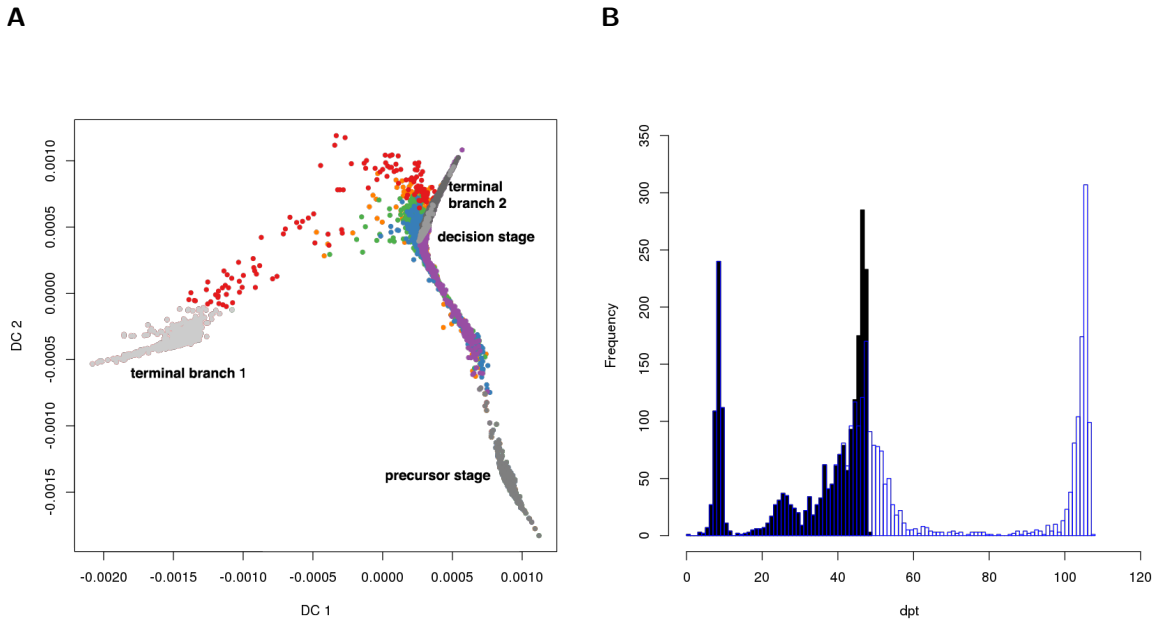
Pseudotime ordering allows to stress the succession of different transcription factors in the qPCR data set. For this purpose, we computed the derivative of the expression along branch 1 and detected the most significant changes. In particular, we used the smoothed version of the data: a sliding window on the gene expression of 50 adjacent cells along the respective branch, where non-detected expression values where modeled with a Gaussian distribution with mean -14 and variance 3 (cf. Fig. 1d in the main text, a non-smoothed version is displayed in Supplementary Fig. 5). Then, we computed an adjusted Z-Score of the expression value with cut-off variance of 3 (in order to prevent largely non-detected genes to increase their noise-level, we used this cut-off in concordance to the noise introduced during smoothing). In addition, the derivative was approximated by a linear regression model over 500 values, largely reducing false positive peaks from noise (see Supplementary Fig. 6).



Supplementary Figure 5: Ordering of genes according to peaks in the smoothed differential of the dynamics shows both switch-like and graded dynamics of the transcription factor cascade. Here, the non-smoothed gene expression is displayed. The colored top bar indicates the embryonic stage of origin for each cell (blue: PS, green: NP, orange: HF, red: 4SG+, purple: 4SG-). The top histogram bar indicates the cell density (high values imply a steady state, low values imply transitions).



Supplementary Figure 6: Determination of switch-like transitions along the pseudotemporal ordering. The first derivative of the pseudotime series is approximately determined by a linear regression coefficient on a sliding window of size 500 along the pseudotime index of the smoothed gene expression. The smoothing is crucial to reduce noise-induced changes of the derivative. Only transitions above the threshold of 0.0028 we considered. A selection of first derivatives is displayed. Notably, there are sharp on- and off-switches (*Kdr*, *Sox17*) and weak or slow transitions (*Notch* on switching and *Etv6* expression).

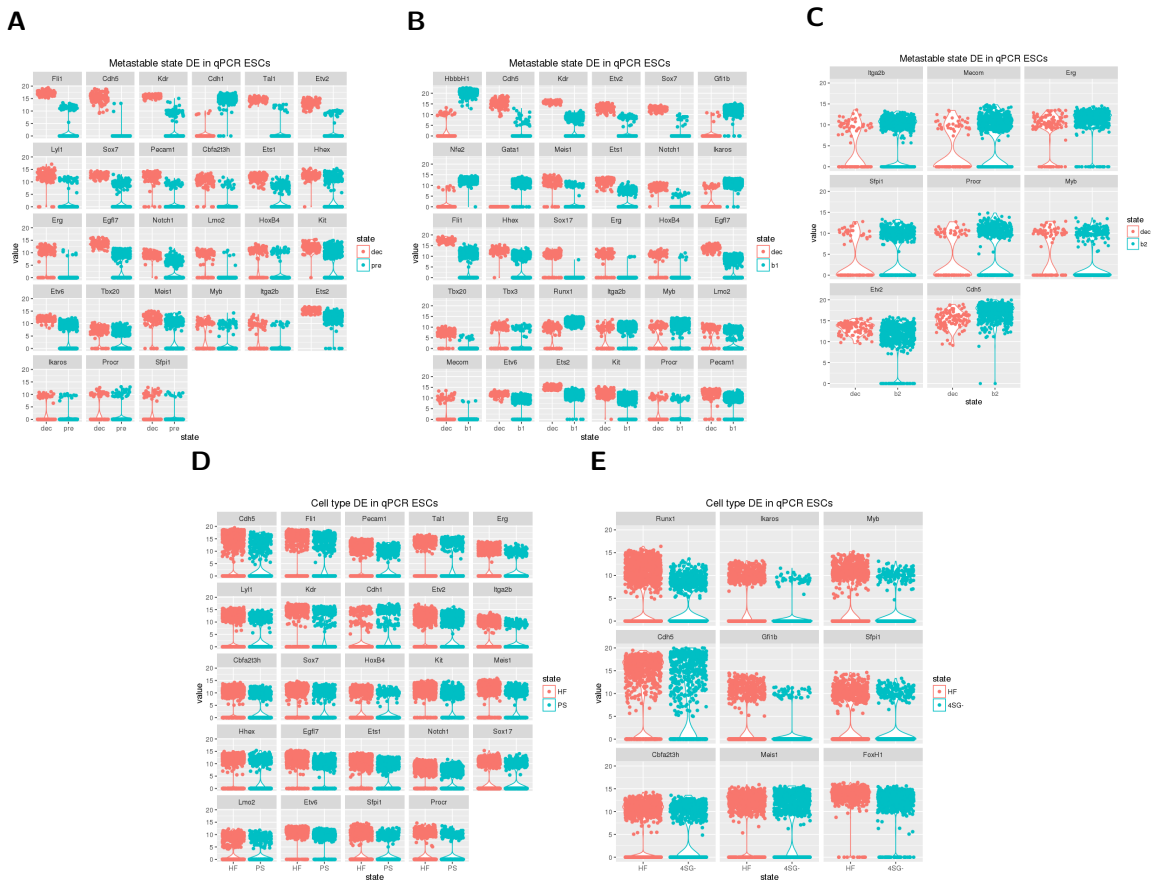


Supplementary Figure 7: A) Diffusion plot illustrating four (metastable) steady states along the pseudotemporal ordering. Lower right: Precursor stage. Left: Terminal branch 1. Upper right: Decision stage (light gray) and terminal branch 2 (dark gray). B) Histogram plot of the cell density along the branches. blue bars: branch 1, black bars: branch 2. Both branches share the precursor branch up to the decision stage.

7.3 Differential expression analysis

We employed the following analysis strategy. First, we introduced metastable states along the pseudotemporal order of the cells (Supplementary Fig. 7), where highly similar cells have approximately the same distance to the root cell. We separated three distinct stages in branch 1 and two stages branch 2 by an appropriate threshold. The decision area is defined by the branch assignment method and is a site of cell accumulation. We highlighted and labeled metastable states in different gray shades (Supplementary Fig. 7A).
Having defined these areas of interest, we fit a modified hurdle model to the gene expression data, calculated the log-fold change of the decision stage to all other states (Supplementary Fig. 8C) and used a likelihood ratio test statistic to compute the significance level.

Furthermore, we repeated the differential expression analysis by the cell types. We compared the gene expression of head fold cells vs. primitive streak cells and 4SG negative cells, respectively. The comparison of head fold and primitive streak is supposed to correspond to the comparison of precursor and decision stage. MAST detected 24 differentially expressed genes (cf. Supplementary Table 1) and most of these genes showed bimodal expression, *Cdh1* has even three levels (cf. Supplementary Fig. 8D). The log fold-change of the detected genes had the sign sign but for *Tbx20*, the absolute values are almost always lower in the cell type comparison than in the metastable state comparison. Comparing the results from the test decision stage vs. terminal branch 2 and head fold vs. 4SG- cells gives a more contradictory picture. The differential analysis of head fold cells and 4SG negative cells detects *Cdh5* as a marker of the endothelial lineage but does not find other markers as *Itga2b*, *Mecom* or *Etv2* that were found in the metastable state setting (decision stage vs. branch 2) (cf. Supplementary Figs. 8C



Supplementary Figure 8: Log-fold change (*lfc*) analysis of the decision area vs. all other states (A-C) and head fold cells vs. primitive streak and 4SG- cells (E,F). The displayed genes are filtered for an $lfc > 2$ and a Bonferroni-adjusted p-value< 0.01. Plots are ordered by absolute *lfc* between the states. A) Decision area (red) vs. Precursor area (blue), B) Decision area (red) vs. branch 1 end point (blue), C) Decision area (red) vs. branch 2 end point (blue). D) Head fold (red) vs. Primitive streak (blue), E) Head fold (red) vs. 4SG negative cells (blue)

and 8E and Supplementary Table 2). Both sets of differential genes share only *Cdh5*, *Myb* and *Sfpi1*. The sign in log fold-change is only consistent in *Myb*, whereas *Cdh5* and *Sfpi1* expression is higher in terminal branch 2 than in the decision stage (positive *lfc*) and lower in 4SG- cells than in head fold cells (negative *lfc*).

	decision stage vs. precursor stage		HF vs. PS	
Gene name	lfc	p_{adj}	lfc	p_{adj}
Cbfa2t3h	-9.77	$1.92 \cdot 10^{-96}$	-4.26	$8.50 \cdot 10^{-61}$
Cdh1	14.21	$1.86 \cdot 10^{-133}$	5.15	$3.05 \cdot 10^{-58}$
Cdh5	-14.86	$8.96 \cdot 10^{-123}$	-8.10	$2.66 \cdot 10^{-119}$
Egfl7	-7.4	$8.36 \cdot 10^{-145}$	-3.19	$4.21 \cdot 10^{-131}$
Erg	-8.21	$2.27 \cdot 10^{-75}$	-5.66	$1.32 \cdot 10^{-112}$
Ets1	-9.52	$5.88 \cdot 10^{-107}$	-3.18	$1.12 \cdot 10^{-57}$
Ets2†	-2.68	$1.07 \cdot 10^{-67}$	-1.89	$3.26 \cdot 10^{-127}$
Etv2	-12.43	$6.68 \cdot 10^{-127}$	-4.81	$5.00 \cdot 10^{-71}$
Etv6	-5.14	$7.11 \cdot 10^{-82}$	-2.36	$1.12 \cdot 10^{-126}$
Fli1	-16.13	$2.44 \cdot 10^{-161}$	-6.99	$6.59 \cdot 10^{-118}$
Hhex	-8.86	$3.58 \cdot 10^{-50}$	-3.41	$1.84 \cdot 10^{-38}$
HoxB4	-6.69	$5.54 \cdot 10^{-42}$	-4.10	$5.27 \cdot 10^{-56}$
Ikaros†	-2.36	$7.19 \cdot 10^{-13}$	-1.40	$2.67 \cdot 10^{-8}$
Itga2b	-4.02	$2.81 \cdot 10^{-25}$	-4.34	$1.07 \cdot 10^{-70}$
Kdr	-14.66	$2.80 \cdot 10^{-163}$	-5.29	$3.77 \cdot 10^{-92}$
Kit	-5.58	$3.60 \cdot 10^{-35}$	-4.08	$1.07 \cdot 10^{-120}$
Lmo2	-6.94	$1.38 \cdot 10^{-69}$	-2.39	$6.10 \cdot 10^{-23}$
Lyl1	-11.46	$2.70 \cdot 10^{-93}$	-5.42	$1.52 \cdot 10^{-78}$
Meis1	-4.40	$3.91 \cdot 10^{-34}$	-4.05	$2.04 \cdot 10^{-122}$
Myb†	-4.05	$9.80 \cdot 10^{-18}$	-1.55	$1.19 \cdot 10^{-7}$
Notch1	-7.17	$1.84 \cdot 10^{-81}$	-3.13	$1.18 \cdot 10^{-73}$
Pecam1	-10.89	$8.51 \cdot 10^{-98}$	-6.35	$8.32 \cdot 10^{-151}$
Procr	-2.28	$9.24 \cdot 10^{-8}$	-2.16	$2.26 \cdot 10^{-17}$
Sfpi1	-2.24	$2.66 \cdot 10^{-11}$	-2.33	$1.59 \cdot 10^{-20}$
Sox7	-10.89	$4.26 \cdot 10^{-121}$	-4.17	$1.05 \cdot 10^{-65}$
Sox17*	-2.02	$1.78 \cdot 10^{-2}$	-2.63	$4.41 \cdot 10^{-24}$
Tal1	-14.04	$1.99 \cdot 10^{-134}$	-5.66	$3.90 \cdot 10^{-76}$
Tbx20†	-4.42	$5.84 \cdot 10^{-30}$	1.51	$7.45 \cdot 10^{-9}$

Supplementary Table 1: Results of the likelihood-ratio test for the gene expression levels in two different stages (decision stage and precursor stage) and cell types (head fold and primitive streak), respectively. We considered a gene expression significant, if the absolute log fold-change was above 2 and the Bonferroni adjusted p-value was below 0.01. † indicates genes being significantly differential in two metastable state comparison but not in cell type comparison. $*$ indicates genes being significantly differential in cell type comparison, but not in metastable state comparison.

	decision stage vs. terminal branch 2		HF vs. 4SG-	
Gene name	lfc	p_{adj}	lfc	p_{adj}
Cbfa2t3h*	0.38	1	-2.50	$4.15 \cdot 10^{-22}$
Cdh5	2.02	$3.06 \cdot 10^{-20}$	-2.66	$1.36 \cdot 10^{-12}$
Erg†	2.85	$5.29 \cdot 10^{-14}$	-1.92	$9.36 \cdot 10^{-21}$
Etv2†	-2.07	$7.59 \cdot 10^{-13}$	-0.45	0.76
FoxH1*	-0.42	$5.83 \cdot 10^{-2}$	-2.17	$2.97 \cdot 10^{-117}$
Gfi1b*	0.37	1	-2.65	$2.35 \cdot 10^{-40}$
Ikaros*	-0.15	1	-4.02	$1.99 \cdot 10^{-87}$
Itga2b†	3.80	$6.86 \cdot 10^{-12}$	-1.5	$3.36 \cdot 10^{-13}$
Mecom †	3.49	$4.12 \cdot 10^{-8}$	0.77	$5.61 \cdot 10^{-15}$
Meis1*	1.65	$3.67 \cdot 10^{-8}$	2.40	$2.06 \cdot 10^{-20}$
Myb	-2.61	$3.55 \cdot 10^{-5}$	-3.29	$5.28 \cdot 10^{-42}$
Procr†	2.64	$2.09 \cdot 10^{-4}$	0.40	1
Runx1*	0.95	1	-4.20	$2.50 \cdot 10^{-94}$
Sfpi1	2.71	$1.62 \cdot 10^{-5}$	-2.65	$1.02 \cdot 10^{-26}$

Supplementary Table 2: Results of the likelihood-ratio test for the gene expression levels in two different stages (decision stage and terminal branch 2) and cell types (head fold and 4SG- cells), respectively. We considered a gene expression significant, if the absolute log fold-change was above 2 and the Bonferroni adjusted p-value was below 0.01. † indicates genes being significantly differential in two metastable state comparison but not in cell type comparison. $*$ indicates genes being significantly differential in cell type comparison, but not in metastable state comparison.

8 DropSeq differentiation data

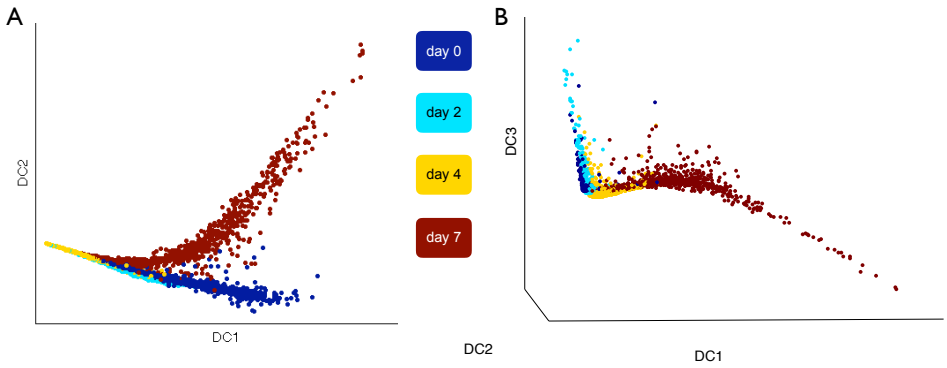
We reanalyzed a single-cell RNA-seq data set using the dropSeq protocol from [12]. Here, single cells along with a set of uniquely barcoded primers were capture in tiny droplets and sequenced. The capabilities of this technique were demonstrated using an undirected differentiation process of mouse embryonic stem cells upon leukemia inhibitory factor (LIF) withdrawal. The data set is available under the GEO accession number GSE65525.

8.1 Pre-processing

There are various sources of variation in single-cell RNA-seq data, beginning with the tiny amounts of RNA molecules to detect to capturing efficiency and amplification bias. The authors in this data set applied both unique molecular identifiers (UMIs) and technical genes to determine a set of 2047 genes with cell-to-cell variance above the technical noise level and normalized by the total amount of transcripts:

$$\hat{m} = m \cdot \frac{\mathrm{E}(M)}{M}, \tag{21}$$

where $M = \sum_i m_i$ is the total amount of UMI-filtered reads m_i per cell and $\mathrm{E}(M)$ is the average of totals over all cells (cf. [12], supplement). We concentrated our analysis on the heterogeneous genes only. A biological source of variance in single-cell transcriptomics is the influence of the cell cycle genes. In particular, differentiating cells are very actively dividing. Recently, [13] introduced the *scLVM* approach to detect the estimate and correct for hidden biological effects as the cell cycle. The method is also capable to reduce batch effects (see Supplementary Figs. 9 and 10A,B). We used the *scLVM* method to account for both technical and cell-cycle induced noise (see Supplementary Fig. 10).



Supplementary Figure 9: A) diffusion map of the raw dropSeq data (before cell-cycle correction). The resolution of cells at day 2 and day 4 is bad because of the cell size in these two days which causes many genes expression to fall below the detection threshold. Such zero terms remain zero even after cell size normalization (i.e. division by sum expression for each cell). B) Cell cycle correction can get rid of the unwanted cell size effect and provides better resolution for the diffusion map.

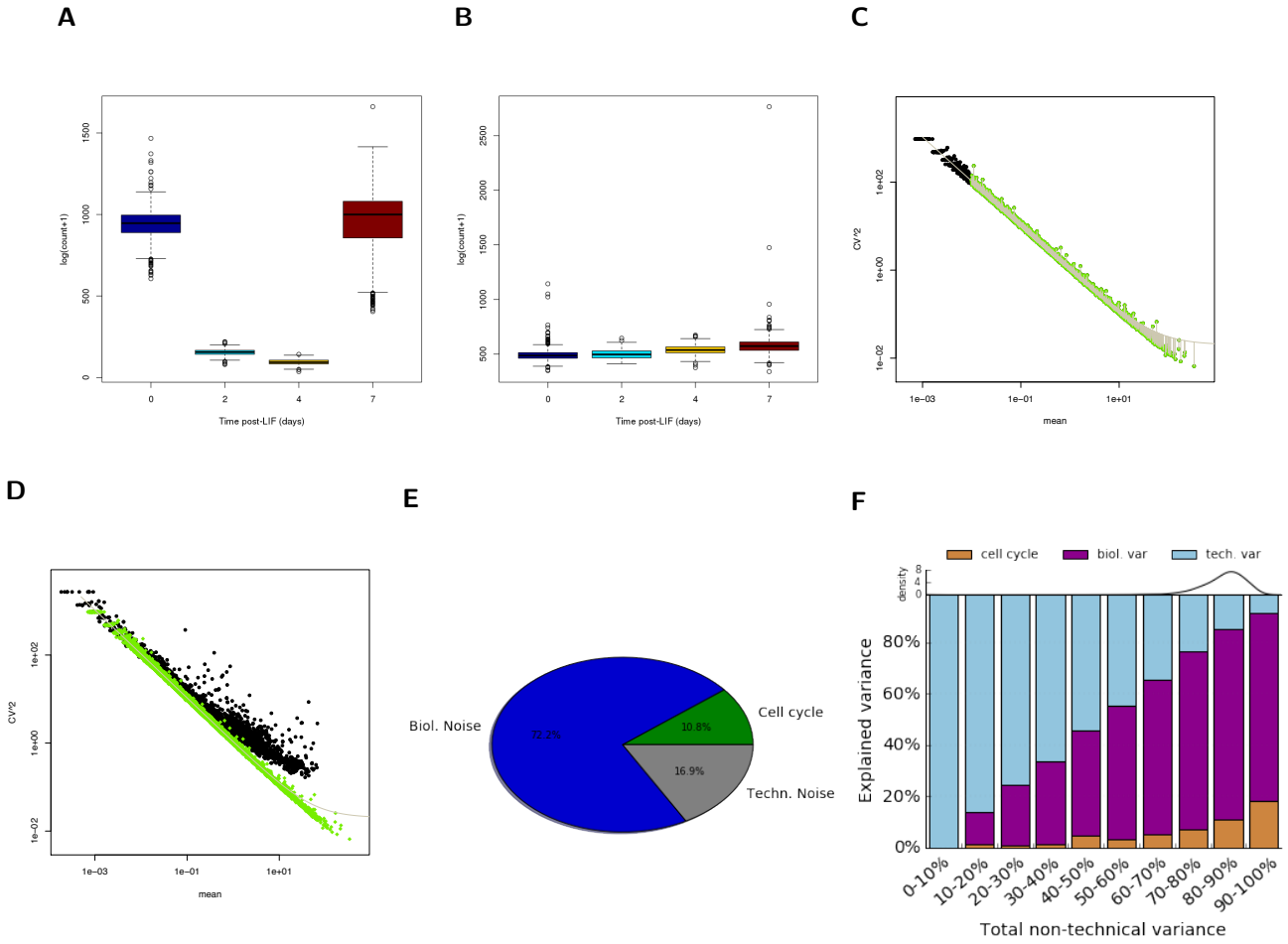
First, we fit the noise model according to Brennecke et al [14] to a pure RNA control sample provided in the data set to estimate the technical noise of the protocol (Supplementary Fig. 10C). For cell-cycle correction, we used the $\log_{10}(\hat{m}+1)$ expression values of the 2044 highly variable genes in 2717 cells measured at 0, 2, 4 and 7 days after LIF withdrawal.

8.2 Clustering

To determine the similarity among genes, computed the gene-to-gene correlation and define $1-\text{cor}$ as a similarity measure. Next, we performed a hierarchical clustering on this similarity measure and highlighted four clusters (cf. Supplementary Fig. 11). We used three different types of normalization to compare the data. First, we have the $\log_{10}(\hat{m}+1)$ normalization we used for the cell-cycle correction. Unfortunately, we observed a very high variance in gene expression after the correction and decided to regularize the cell-cycle corrected data.
First, we used a quantile normalization for the expression y_{ig} as follows: We compute the 0.02- and 0.98-percentiles $(p_{g,0.02}, p_{g,0.98})$ for each gene g and calculate

$$\tilde{y}_{ig} = \frac{y_{ig} - p_{g,0.02}}{p_{g,0.98} - p_{g,0.02}} \tag{22}$$

Then, all expression values within the $[p_{0.02}, p_{0.98}]$-interval are normalized to the $[0,1]$-interval and outliers are found outside this interval.
Second, we applied a Z-Score-normalization with zero mean and unit variance for the expression y_{ig}.



Supplementary Figure 10: Cell-cycle correction of mESC dropSeq data using *scLVM*. A,B) The total count of transcripts from 2044 heterogeneous genes per day. A) log-normalized counts before cell-cycle correction. B) log-normalized counts after cell-cycle correction. C) Fit the CV2-mean relation according to Brennecke et al [14] to a pure RNA control and D) superimpose these technical genes with endogenous genes. E) Variance decomposition according to the identified latent variables. F) Detailed variance decomposition sorted by technical noise contribution.

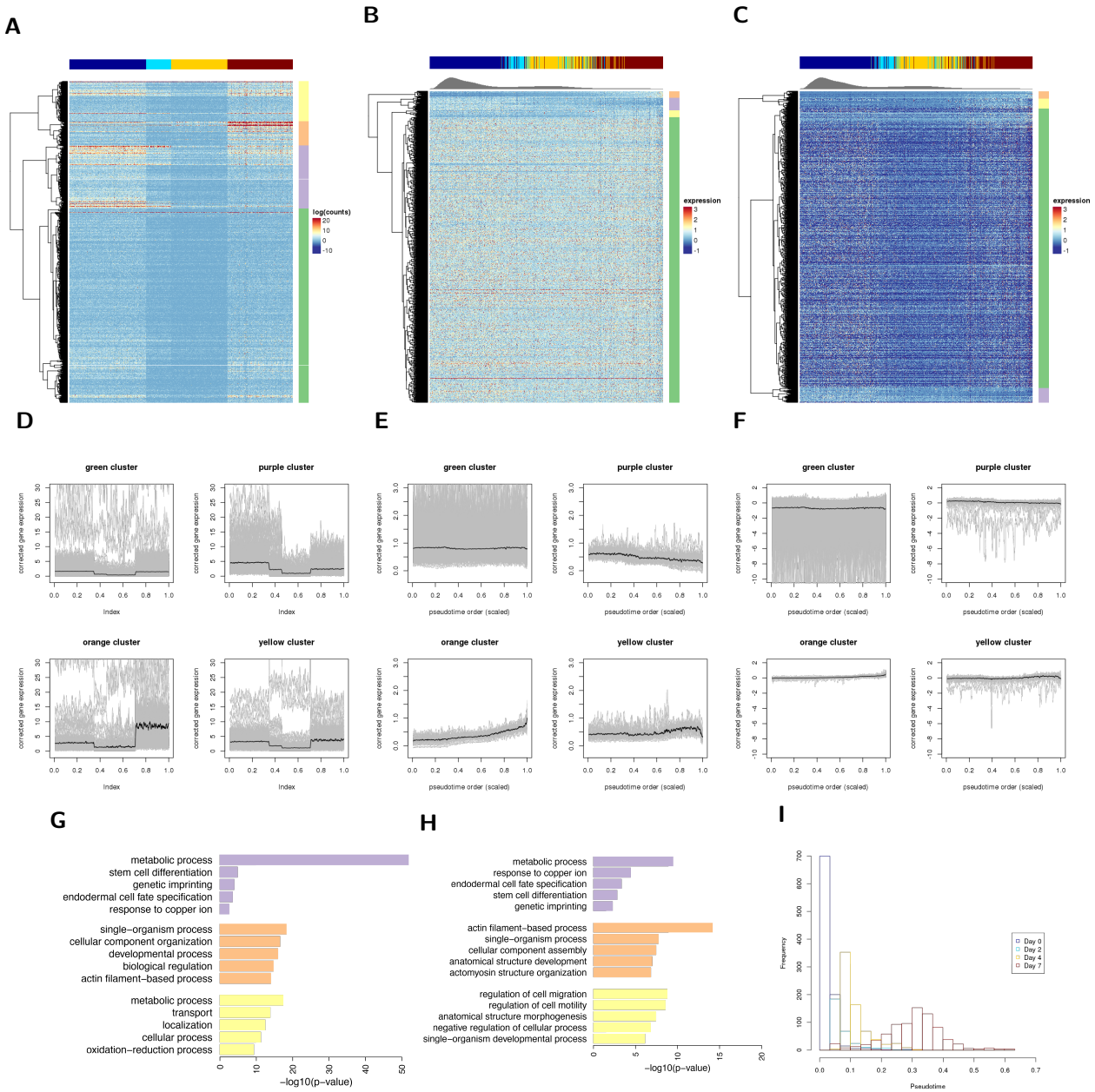
Both normalizations regularize the cell-cycle corrected gene expression. Though, the quantile normalization is more robust to outliers and alleviates their detection. First, batch and cell cycle correction decreased the day-to-day variability of the samples and by pseudotemporal ordering, the differentiation process was resolved in greater detail. We are able to spot a single differentiation path in this data set as well as different subpopulations (cf. Supplementary Figs. 10A,B and 11). The diffusion pseudotime embedding resolves the heterogeneity of the measurement days (top annotation in Supplementary Fig. 11A-C and Supplementary Fig. 11I). As we consider pseudotime as a measure of differentiation in this case, small pseudotimes correspond to a low degree of differentiation and a high degree of pluripotency, respectively. We observe an increasing degree of heterogeneity with time passed since LIF withdrawal and as reported in [12], we observe large variability at day 7 ranging from pluripotent cells to strongly differentiating cells. A detailed interpretation regarding gene expression patterns is conducted upon gene clusters.

8.3 GO enrichment analysis

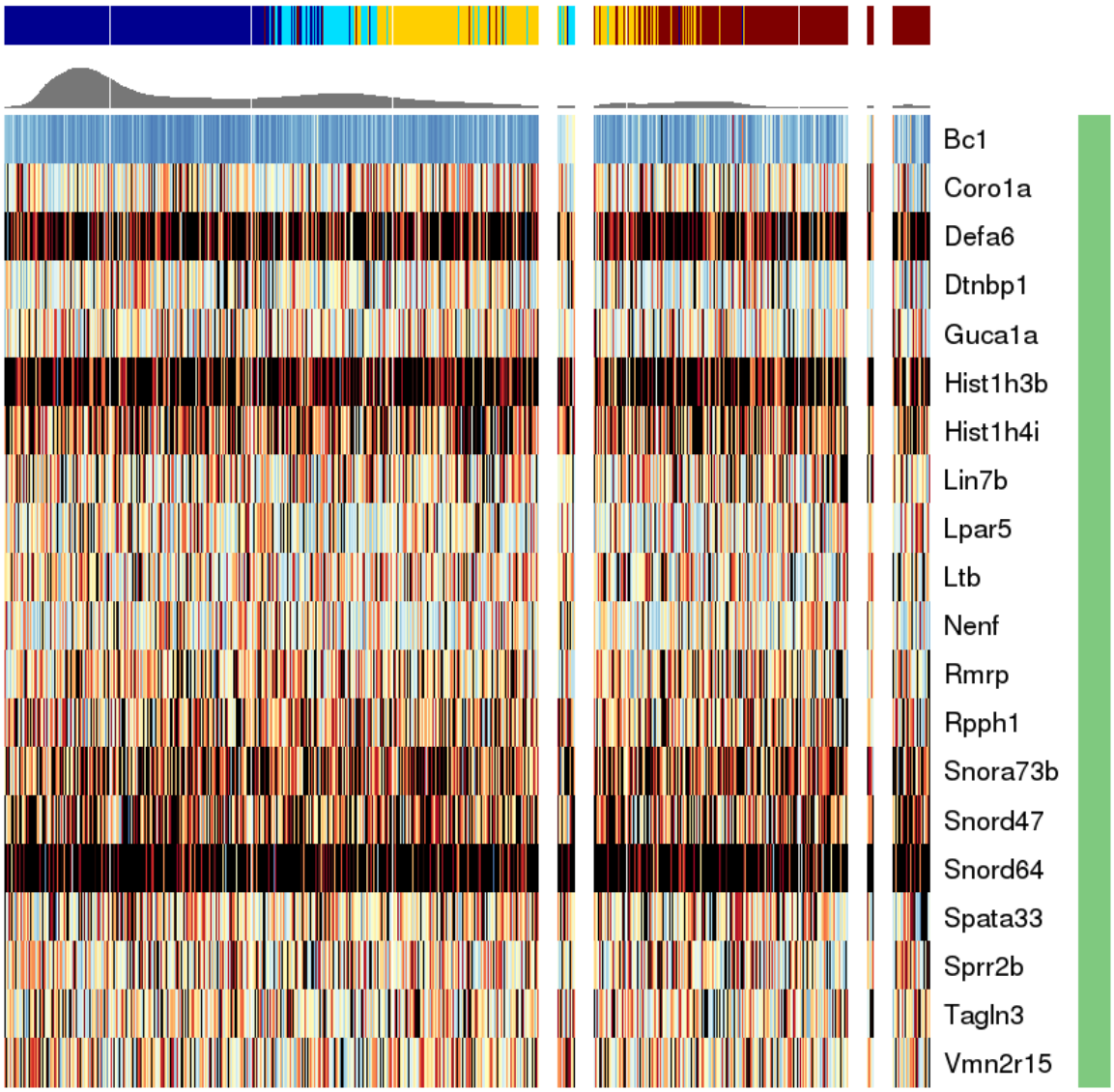
To assess the reasonability of the hierarchical clustering, we performed a GO enrichment analysis using Genomatix software suite (*www.genomatix.de*) before and after cell-cycle correction (Supplementary Fig. 11G,H). For illustration, we picked five GO terms for each cluster. Clustering of the non-cell-cycle corrected data revealed strong differences compared to the GO enrichment after cell-cycle correction. Indeed, the GO terms of the Z-score and quantile normalized data are very similar in all clusters. To demonstrate the differences arising with cell-cycle and batch correction, we compared GO terms given in the purple cluster (Supplementary Fig. 11G). The p-value of GO:0008152 (metabolic process) is $1.37 \cdot 10^{-52}$ indicating a strong metabolic signature. The other displayed terms range among the top 400 GO terms in this list where we also have key factors of pluripotency promoting endodermal cell fate specification (GO:0001714) (*Pou5f1, Sox2, Nanog*) with a p-value $3.5 \cdot 10^{-4}$. However, in the yellow cluster is also a strong metabolic signature (p-value $4.08 \cdot 10^{-18}$), but deviates from the terms in the cell-cycle corrected data as we do not find GO enrichment for cell migration (GO:0016477) or regulation of cell motility (GO:2000145). Hence, in order to recover cellular events with GO enrichment, we need to consider a robust data normalization.

8.4 Differential expression on branches

We found several branches corresponding to subpopulations of differentiating ES cells. We found a first population branching off mainly consisting of day 2 cells and at the late stage another split into epiblast-like and primitive endoderm-like cells. To test which genes are differentially expressed in the first side branch, we performed rank sums tests of 250 cells from early stage, the side branch cells and the epiblast-like cells, each. We only considered those genes, that have approximately the same expression level in early and epiblast-like stage, but differed from both stages in the side branch cells (p-values Bonferroni adjusted). All p-values < 0.01 were considered (cf. Supplementary Fig. 12). We identified 20 genes with a strong variation along the pseudotemporal order as they all belong to the green cluster (cf. Supplementary Figs. 11).



Supplementary Figure 11: Heatmap displaying the expression profiles of 2044 highly variable genes before A) and after cell-cycle correction and pseudotime ordering (B,C), time courses of gene expression along batch (D) and pseudotime (E,F), GO enrichment analysis of the clusters in (A,C). The colored top bar (A-C) indicates the time after LIF withdrawal (dark blue: day 0, light blue: day 2, yellow: day 4, red: day 7). A) Gene expression with strong day-to-day variability. B) Cell-cycle corrected gene expression and additional quantile normalization. C) Cell-cycle corrected gene expression and additional Z-score normalization. Pseudotemporal ordering is indicated by mixed colors in the top annotation bar. In the time courses, the respective genes are indicated in grey, the black curve is the smoothed mean. D) log-transformed gene expression counts. E) Cell cycle correction, log transformed gene expression counts, quantile normalization (cf. Fig. 2d in main text). F) As in E), with Z-score normalization. All clusters share the same temporal behavior. The green cluster GO terms are not shown. For each cluster, five representative GO terms are displayed. G) GO terms before cell-cycle correction, H) after cell-cycle correction and Z-score normalization. I) Distribution of cells along pseudotime labeled by time after LIF withdrawal.



Supplementary Figure 12: Heatmap with differentially expressed genes in the first side branch. Expression values are quantile normalized (gray - lowest $2\,\%$, blue - red: low to high expression, black - highest $2\,\%$.

9 Comparison with Wanderlust and Monocle

We compared the performance of DPT with Monocle and Wanderlust on several data sets;

- the human myoblast cells RNA-Seq data [15],
- mESC dropSeq data [12],
- early blood cells qPCR [11], and
- human B-cells sampled from four different patients measured by mass cytometry [4].

Except the latter data set ([4]), for which time labels where not available, we performed pseudotime ordering of bootstrap sets of about 70% of the total number of cells for each data set with all three methods and compared the concordance of each method's pseudotime order with that of DPT as described in section 9.1. Because Monocle fails to run on large number of cells, when necessary, a smaller sample size (700 cells) was used for it in the bootstrap sets. Supplementary Table 3 and Figure 2C in the main text summarize the results of bootstrap comparisons. DPT generally shows a higher mean and smaller variance of concordance with time labels compared to the other methods.

9.1 Concordance of pseudotime ordering with time labels.

The concordance for each subsample was measured as Kendall tau correlation of each pseudotime order with the time labels of that subset. We then calculated the mean (μ_1) and variance σ_1^2 of the concordance measure over several bootstrap sets. A 2-sided t-test was performed to specify the significance of difference in performance compared to the diffusion pseudotime (DPT) ordering, specified by index 2 in the following calculation:

$$\mathrm{t} = \sqrt{\frac{m_1 \cdot m_2}{m_1 + m_2}} \cdot \frac{|\mu_1 - \mu_2|}{s} \tag{23}$$

where m_1 denotes the number of bootstrap runs with the first method and m_2 denotes the number of bootstrap runs with DPT. The weighted variance s is computed as

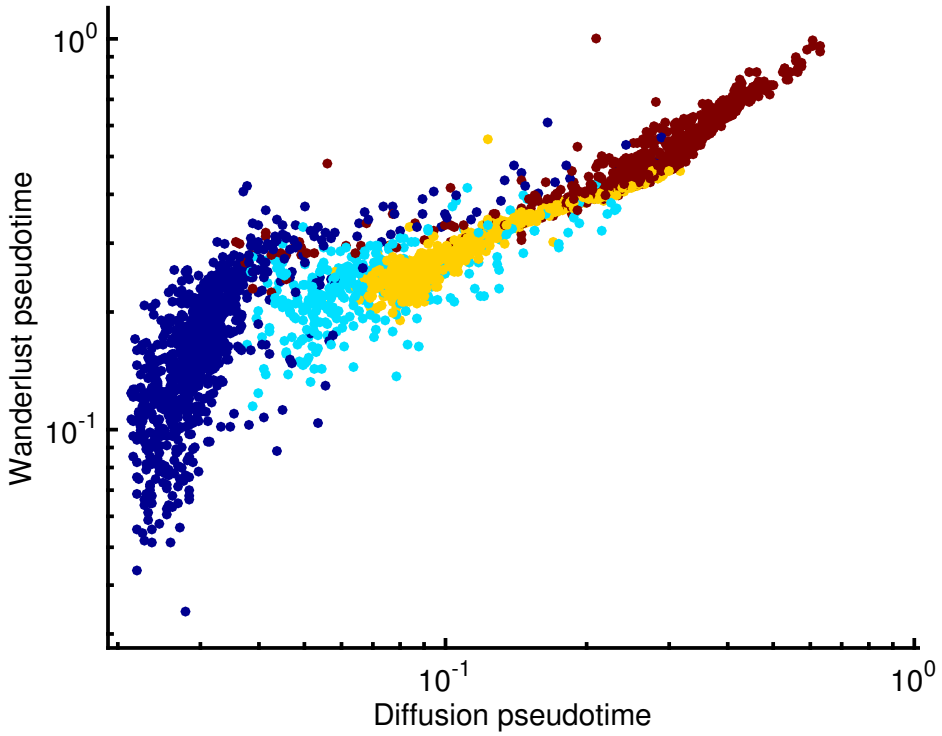
$$s = \sqrt{\frac{(m_1 - 1)\sigma_1^2 + (m_2 - 1)\sigma_2^2}{m_1 + m_2 - 2}}, \tag{24}$$

where $m_1 + m_2 - 2$ equals the degrees of freedom (df). The p-values were then computed using the tcdf function in Matlab as $p = 2 \cdot (1 - \mathrm{tcdf}(t, \mathrm{df}))$.

9.2 Visual comparison for human myoblast cells RNA-Seq and human B-cells mass cytometry data sets.

The performance of the three methods on the complete data (i.e. not bootstrapped) of human myoblast cells ([15]), are visually compared in Supplementary Figure 14. We chose to plot the expression of five genes with known functions (see [15]) over the pseudotime of each method.

We also compared the performance of Wanderlust and DPT also on six data sets of human Bcells (sampled from six different patients) measured by mass cytometry [4]. Again Monocle fails to perform on such large body of cells ($n > 10^3$). This data did not have any time labels thus DPT and Wanderlust pseudotimes colored on the diffusion map manifold, as well as three markers (CD45, IgM, TdT) expression over each pseudotime is presented in Supplementary Figure 15 for visual comparison.



Supplementary Figure 13: Wanderlust pseudotime against diffusion pseudotime for the mESC dropSeq data set.

9.3 Comparison of DPT and Wanderlust on dropSeq mESC cells

For mESC dropSeq data [12], pseudotime distribution of cells in the experiments from the four different days, for DPT and Wanderlust were computed. Diffusion pseudotime orders cells well along the four temporal categories (Kendall rank correlation 0.784 ± 10^{-5}), significantly better than pseudotemporal ordering by Wanderlust (Kendall rank correlation 0.716 ± 10^{-5}). In Supplementary Figure 13 we plot the Wanderlust pseudotime against DPT.

9.4 Discussion on the differences between DPT, Wanderlust and Monocle

In this section we discuss the performance of the three methods (Wanderlust, Monocle and DPT) in respect to several challenges in single-cell pseudotime ordering. Supplementary Table 4 provides an overview of this comparison.

9.4.1 Methodology

Here we give a simplified and brief explanation of the Monocle and Wanderlust methods. Monocle orders cells on the Minimum Spanning Tree built on a few Independent Component Analysis embedding dimensions. Wanderlust first builds nearest neighbors graph on cells. Assuming a nonbranching manifold, Wanderlust can allow treating relative distances as a scalar (sometimes termed "displacement"), where addition or reduction of the displacements (i.e. orientation of the 1D manifold with respect to the root cell) is decided by a number of (random) landmarks on the data. It then averages the displacement of each cell relative to the root cell for a finite number of sampled paths on the graph. The methodology for DPT is described in detail in section 1.1.

	human myoblast (RNA-seq)	early blood cells (qPCR)	mESC (RNA-seq/dropSeq)
Monocle	$n' = 190$, $m = 100$ $\tau = 0.413 \pm 0.187$ $p = 0.56$	$n' = 700$, $m = 100$ $\tau = 0.356 \pm 0.024$ $p = 0$	$n' = 700$, $m = 100$ $\tau = 0.446 \pm 0.141$ $p = 0$
Wanderlust	$n' = 190$, $m = 100$ $\tau = 0.407 \pm 0.054$ $p = 9.5 \cdot 10^{-3}$	$n' = 2700$, $m = 100$ $\tau = 0.381 \pm 0.01$ $p = 1.2 \cdot 10^{-4}$	$n' = 1800$, $m = 100$ $\tau = 0.667 \pm 0.022$ $p = 0$
DPT	$n' = 190$, $m = 100$ $\tau = 0.424 \pm 0.036$	$n' = 2700$, $m = 100$ $\tau = 0.386 \pm 0.008$	$n' = 1800$, $m = 100$ $\tau = 0.782 \pm 0.004$
original data size after pre-processing ($n \cdot G$)	$271 \cdot 1548$	$3934 \cdot 42$	$2771 \cdot 2044$

Supplementary Table 3: Comparison of Monocle, Wanderlust and DPT on concordance to the time labels over three different data sets. In all cases DPT shows a higher mean and smaller variance of concordance to the time labels compared to the other methods. n denotes the number of samples, n' the number of samples in the subset used for analysis, G the number of genes and m the number of bootstrap runs. τ is the result of the Kendall rank correlation (p-values depicts significance of Monocle and Wanderlust to DPT, respectively.)

9.4.2 Definition of pseudotime

Monocle notes pseudotime as distances on a single possible path (on MST) on an ICA dimension reduced embedding of data.

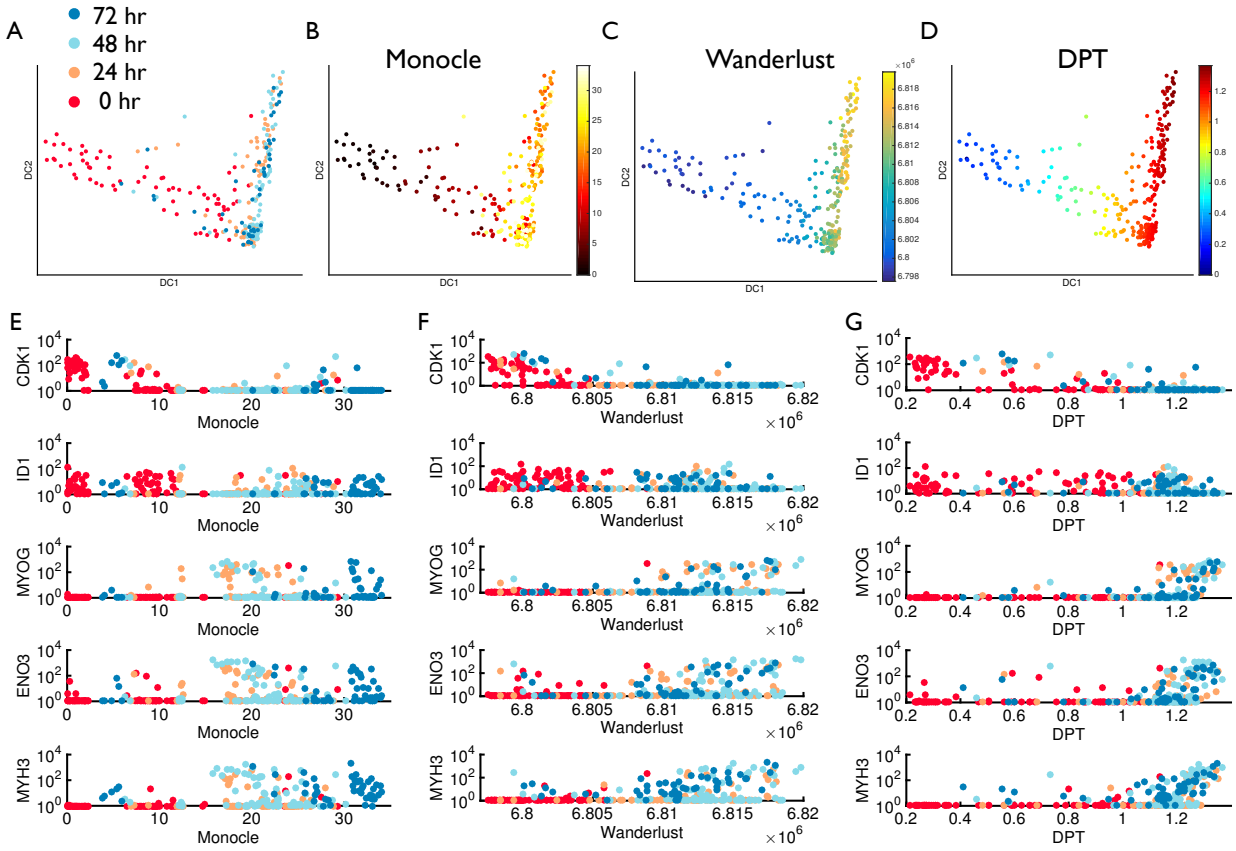
If performed on data in the original gene expression space ($\mathbb{R}^G$), Wanderlust aims to recover the geodesic distances from the root cell on the original manifold $C \subset \mathbb{R}^G$, which would exactly be the same as universal time (see Supplementary Figs. 2 and 3C in section 1.1). However, Wanderlust's path sampling approach hinders the reliable recovery of such geodesics for more complicated manifold structures (e.g. sharp turns, branching) and in presence of large noise in the data. It goes without saying that when performed on (mapped) reduced dimensions, Wanderlust pseudotime cannot be taken as a notion of universal time anymore.

DPT uses a path integral approach to account for all possible paths between cells. Diffusion pseudotime is finally defined as Euclidean distance to the root in a mapped $\mathbb{R}^{(n-1)}$ space with $\frac{\lambda_i}{1-\lambda_i}\psi_i, i = 1, ..., n-1$ coordinates.

9.4.3 Robustness to noise and subsampling

ICA (as a linear embedding method) can not catch the non-linearity of differentiation manifolds present in many differentiation systems. When ICA is replaced by non-linear method such as diffusion maps, monocle usually shows better performance (put some demonstration figures). On the next level monocle's pseudotime order is based on only one single connecting path which is the minimum spanning tree. Thus monocle does not consider the possibility of reaching to a state through multiple paths, which is in principle possible in differentiation systems. Furthermore the single path might provide a reasonable pseudotime order for clean data, but as soon as there is bit higher noise in the data, biologically non-relevant short cuts come into play and biologically relevant information of alternative paths get neglected all because the pseudotime measure based on a single path is not robust to noise.

Wanderlust provides considerable robustness to noise by sampling multiple paths to connect each cell to the initial cell (rather than choosing only one single path). However the path sampling approach is dependent on sampling density of several cell subpopulations and the algorithm lacks any correction for



Supplementary Figure 14: Diffusion map embedding for human myoblast differentiating cells with colorcode corresponding to A) time labels. B) Monocle pseudotime. C) Wanderlust pseudotime D) DPT. Expression for a few genes over E) Monocle pseudotime, F) Wanderlust pseudotime, G) Diffusion pseudotime.

density heterogeneity effects on the paths sampling. This can explain the larger variance of Wanderlusts' concordance with time labels compared to DPT (see Supplementary Table 3 and Figure 2e in the main text).

DPT's path integral (considering all possible paths) approach renders it quite robust to noise. Furthermore in the implementation of the diffusion map (used by DPT) we correct for density heterogeneity effects [6].

9.4.4 Dealing with stationary (metastable) cell states

There is barely any biological meaning to ordering cells in the metastable states. Although in actual time dynamics there is such an order even in the metastable states for a single cell trajectory, the order in a metastable state can be completely shuffled in another differentiating single cell's trajectory. That is, all the states belonging to the same metastable state will have almost the same universal time value. Thus We suggest that pseudotime should be measured as the distance to the initial cell on a reconstruction of differentiation manifold. That is no matter how long in actual time a single cell trajectory is trapped in one of the metastable states, there is almost no progression on neither the original manifold ($C \subset \mathbb{R}^G$) or any mapping of it (C'). Hence one should expect almost the same

pseudotime for cells in a metastable state.

Monocle however provides an absolute order of cells based on the constructed MST, which is not reliable at the metastable states.

For nonbranching data, Wanderlust provides a valid pseudotime measured on the manifold by adding up Euclidean distances of neighboring cells for a sampled path that connect each cell to the initial cell.

DPT measures pseudo-time on a mapped manifold such that cells from the same metastable state are correctly places on the same neighborhood on the manifold and a similar distance relative to the root cell is assigned to them.

9.4.5 Applicability to large cell numbers and run time

Monocle fails to find the MST solution (generates an error message) when applied on a large body of cells ($> 10^3$) that dont show much structure in the ICA embedding. This tends to happen more often if a very large set of genes are used for ordering. ICA runs in computation time of order Gn and MST's run time scales with the number of edges, thus n^2. Thus the total computation time for monocle is $n(G + n)$.

With appropriate choice of the tuning parameters Wanderlust preforms well for large cell numbers. The computational time for finding nearest neighbors is $\log(n)$. Thus wanderlust performs in $\log(n) + n$num.landmarks.

The computational time needed for DPT using the complete transition matrix is $O(Gn^2)$. However the transition matrix T is sparse, most entries are close to zero. Two approximations are possible: (i) enforcing a threshold from which entries are set to zero. (ii) a k nearest neighbor approximation. Both leads to a block tridiagonal matrix. In the simplest case, this is a tridiagonal matrix which can be inverted at the cost of O(n). In the case of a band matrix with $k/2$ diagonals next to the main diagonal, one can compute the crucial step of matrix inversion - a LU factorization - in $O(nk^2)$ [16]. Further steps only require $O(nk)$. In the completely general case with a varying number of non-zero entries k_i per row, one can make use of algorithms with similar complexity [17].

9.4.6 Allowing for multiple root cells

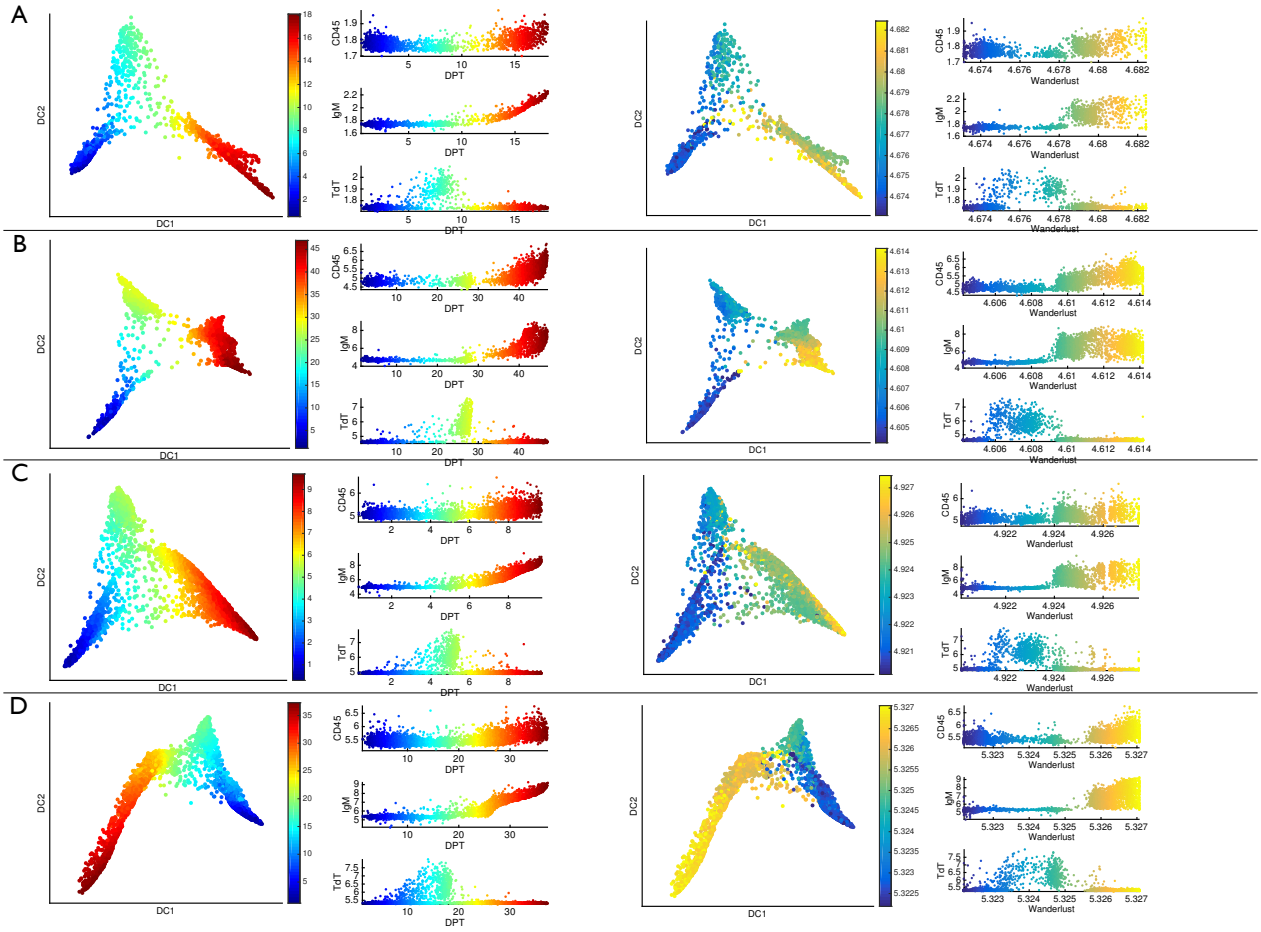
The heterogeneity of differentiating cell populations can also hold for the pluripotent state. Very often all the pluripotent cells reside on a close neighborhood on the differentiation manifold. In this case even with a single chosen root cell, the pseudotime measured on the manifold would automatically account for the metastable state of pluripotency. However one could expect larger variability among the pluripotent cells or even existence of several sources for pluripotent cells which could consequently take different probabilities for paths of development depending on the state of the root cell. DPT thus provides the possibility of choosing multiple root cells. Neither monocle and Wanderlust provide this feature.

9.4.7 Data embedding and visualization

Monocle visualizes the ICA embedding of data.

Wanderlust assumes a one dimensional (i.e. nonbranching) order of cells and does not provide any specific visualization of the data manifold.

DPT's mapped manifold's space shares the same eigenvectors with diffusion maps except for the non-informative (zeroth) eigenvector which is left out in diffusion map embedding as well. Thus diffusion map embedding is a consistent visualization for DPT.



Supplementary Figure 15: Comparison of DPT (left in jet colormap) performance to Wanderlust (right in parula colormap) visualized on diffusion map for four samples (A to D) of human B-cells and the expression of three genes versus each pseudotime.

algorithm	reference	methodology	tuning parameters	computation time	handles branching lineages	scalability to large sample numbers	robustness to noise/ density heterogeneity	allows several root (pluripotent) cells	pseudotime is measured on full data dimensions	provides data embedding	handles missing/ uncertain values
Monocle	Trapnell et al[15]	MST on ICA embedding	number of ICA embedding dimensions, root cell	$\mathcal{O}(n^2 + G \cdot n)$	+	−	−	−	−	+	−
Wanderlust	Bendall et al[4]	sampling paths on nearest neighbors graphs	k, l, num_graphs (n_g), num_landmarks (n_l), root cell	$\mathcal{O}(n \cdot n_l)$	−	+	+	−	+	−	−
DPT		analytic path integral	diffusion parameter (σ or κ), root cell	$\mathcal{O}(n \cdot k)$	+	+	+	+	+	+	+

Supplementary Table 4: Comparison of several single-cell pseudotime ordering algorithms.

References

[1] K. Campbell, C. P. Ponting, and C. Webber, "Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles," *bioRxiv doi: 10.1101/027219*, 9 2015.

[2] K. Campbell, C. Yau, C. P. Ponting, and C. Webber, "Bayesian gaussian process latent variable models for pseudotime inference in single-cell RNA-seq data," *bioRxiv doi: 10.1101/026872*, 9 2015.

[3] R. Kafri, J. Levy, M. B. Ginzberg, S. Oh, G. Lahav, and M. W. Kirschner, "Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle.," *Nature*, vol. 494, no. 7438, pp. 480–483, 2013.

[4] S. C. Bendall, K. L. Davis, E.-A. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er, "Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.," *Cell*, vol. 157, no. 3, pp. 714–725, 2014.

[5] G. Gut, M. D. Tadmor, D. Pe'er, L. Pelkmans, and P. Liberali, "Trajectories of cell-cycle progression from fixed cell populations.," *Nature methods*, vol. 12, no. 10, pp. 951–954, 2015.

[6] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, pp. 2989–98, 5 2015.

[7] C. D. Meyer, Jr., "The role of the group generalized inverse in the theory of finite markov chains," *SIAM Review*, vol. 17, pp. 443–464, 7 1975.

[8] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[9] F. Fouss, A. Pirotte, and M. Saerens, "The application of new concepts of dissimilarities between nodes of a graph to collaborative filtering," 1 2004.

[10] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo, "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data," *Genome Biology*, vol. 16, no. 1, p. 278, 2015.

[11] V. Moignard, S. Woodhouse, L. Haghverdi, A. J. Lilly, Y. Tanaka, A. C. Wilkinson, F. Buettner, I. C. Macaulay, W. Jawaid, E. Diamanti, S.-I. Nishikawa, N. Piterman, V. Kouskoff, F. J. Theis, J. Fisher, and B. Göttgens, "Decoding the regulatory network of early blood development from single-cell gene expression measurements," *Nature Biotechnology*, vol. 33, no. 3, pp. 269–76, 2015.

[12] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, pp. 1187–1201, 5 2015.

[13] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, pp. 155–160, 1 2015.

[14] P. Brennecke, S. Anders, J. K. Kim, A. A. Koodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler, "Accounting for technical noise in single-cell RNA-seq experiments.," *Nature methods*, vol. 10, pp. 1093–5, 11 2013.

[15] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.," *Nature biotechnology*, vol. 32, pp. 381–6, 4 2014.

[16] E. Klç and P. Stanica, "The inverse of banded matrices," *Journal of Computational and Applied Mathematics*, vol. 237, no. 1, pp. 126–135, 2013.

[17] D. E. Petersen, H. H. B. Sørensen, P. C. Hansen, S. Skelboe, and K. Stokbro, "Block tridiagonal matrix inversion and fast transmission calculations," *Journal of Computational Physics*, vol. 227, no. 6, pp. 3174–3190, 2008.